Resources

Gathering the 'Net: Efforts and Challenges in Archiving Pacific Websites

ELEANOR KLEIBER

The Contemporary Pacific, Volume 26, Number 1, 157–166

Gathering the 'Net: Efforts and Challenges in Archiving Pacific Websites

Eleanor Kleiber

In addition to more traditional material—books, journals and other serial publications, brochures, music, films, manuscripts, photographs, postcards and archives—the University of Hawai'i–Mānoa (UHM) Library's Hawaiian and Pacific Collections are now actively collecting websites. With so many new websites being created in and about the Pacific Islands region, and so much more information being made available online—and at times exclusively so—it has become increasingly clear to the librarians of these collections that to adequately document this period in history it is necessary to collect and preserve websites. The UHM Library has been attempting to archive websites in one form or another since 2001. This essay will discuss the importance of collecting Pacific websites, describe how the Hawaiian and Pacific Collections are finding solutions for the inherent challenges of preserving websites, and explore some potential future directions that would strengthen the project and meet the information and research needs of the Pacific Islands region.

Why Preserve Websites

The UHM Library's Hawaiian and Pacific Collections have long recognized the research value of websites. In many ways the UH effort seeks to parallel a web archiving project being conducted at the Library of Congress (LOC), which (as noted on the LOC website) is "composed of sites selected by subject specialists to represent web-based information on a designated topic. It is part of a continuing effort by the Library to evaluate, select, collect, catalog, provide access to, and preserve digital materials for researchers today and in the future."[1] In the most recent survey of web archiving initiatives, forty-two institutions were identified world-

The Contemporary Pacific, Volume 26, Number 1, 158–166

wide, with the vast majority of those focused on archiving websites limited to the hosting institution, its nation, or its region or state within a nation (Gomes, Mirana, and Costa 2011). This same survey identified the UHM web archiving project as one of only two initiatives that seek to collect internationally, based on a specific geographic area[2]—although, as will be discussed, the national libraries of Australia and New Zealand do collect beyond their national boundaries.

Vast amounts of information from and about the Pacific are being created digitally and made available via the Internet. Governments, nongovernmental organizations, and businesses use web pages as informational brochures, as tools for interacting with and serving citizens and customers, and as repositories for their work. Blogs are a tool frequently used by citizens to voice individual opinions, at times counter to government preference. Pacific websites are also a meeting place for the Pacific diaspora, connecting far-flung communities and collecting their voices as new cultural identities are navigated and shaped. And, of course, Pacific websites create strong connections between the Pacific region and the rest of the world. Pacific history, culture, and identity are being created and explored using the medium of the web, and librarians in the UHM Hawaiian and Pacific Collections feel it is important that all of this information is preserved for the long-term benefit of researchers and for the citizens of the Pacific.

Although of extreme importance, websites are not a stable medium and thus cause multiple challenges for long-term preservation. For one, websites are constantly changing and being updated, much as if there was a new edition of the same book being published daily. Therefore the same website must be repeatedly "collected." Organizations that produce serial publications may only maintain the most current issue online, assuming that older issues are not of interest (when in fact they can be of great historical value). Then there is the regularity with which entire websites completely and irretrievably disappear. Broken links and "404 not found" messages are well-known and frustrating realities to even the most casual user of the Internet. While printed annual reports do not disintegrate when organizations shut down, websites disappear as soon as there is no funding to maintain them on a server. Government websites are highly susceptible to change or deletion in response to shifting political agendas. With so many economic and social fluctuations that have immediate impact on the existence of a website, libraries should not and cannot rely on the creators of web content to also act as the preservers of that content.

Libraries have a responsibility to serve as a stabilizing force to ensure the preservation of "born digital," Internet-based material, just as they have done for centuries with physical material.

Another reason that the UHM Pacific Collection has prioritized archiving of the Pacific web is that the libraries and archives of most of the Pacific Island countries and territories do not yet have the legal mandate or the capacity to engage in web archiving.[3] In fact, besides UHM Pacific Collection, there are only two other institutions that archive the Pacific web: the National Library of New Zealand and the National Library of Australia, each focusing on those islands with which they have strong historic and political links. And although the UHM Pacific Collection itself has no legal mandate to collect Internet-based material, it has what might be termed a moral imperative: Since its inception in the 1960s, Pacific Collection librarians have seen it as their duty to comprehensively collect and preserve a full and complete record of life in the Pacific. Preserving websites is an extension of this work and is vital to the goal of documenting the Pacific experience as it exists in the early stages of the twenty-first century.

How We Preserve Websites

In some cases, the shift from print to web-based material is relatively simple. There are frequent instances where websites serve as the access point to digital versions of published materials such as books, annual reports, and newsletters. At times, parallel print and online versions are disseminated, such as the Australian National University's State, Society and Governance in Melanesia (SSGM) Discussion Papers.[4] An ever-increasing number of publications are available exclusively through a website or through an online subscription, such as Pitcairn Island's Pitcairn Miscellany.[5] In the Pacific region, the shift to electronic-only publishing is also being sped up by countries choosing to phase out their government printing offices, which in turn leaves publication up to individual agencies and departments. With little or no funding to produce print copies, web-based publishing has become the preferred method of distribution. In such cases, the current policy of the UHM Pacific Collection is to download these documents, print them out, and then have them bound, cataloged, and added to our physical collections. With these procedures, the UHM Pacific Collection essentially functions as a regional government printing office. As more material is published exclusively online, it is recognized that this

practice is not sustainable and so we investigate alternatives such as digital archiving. The long-term storage of the item might change depending on whether we print it or choose to preserve it electronically, but the traditional forms of bibliographic description and access via the UH Voyager online catalog are maintained.

But most websites are much more than repositories for digital documents. The increased sophistication of websites requires preserving not only content but also context. Even the more basic websites have links and navigation that are difficult to adequately convert to print. In 2001, when the Hawaiian and Pacific Collections were first attempting to address the question of how to preserve websites, the solution was to print them out. Not only was this not sustainable due to paper, labor, and space limitations, it was also not a satisfying solution because the context of the website was lost in the translation to the printed page. While the context of a physical book is determined by its binding and (most often) limited to a linear and chronologically paginated system, the context of a website is (not surprisingly) much more like a web. That interlocking context is lost with the physical printout. It was thus necessary to preserve websites in their native digital format with all the sophisticated functions intact.

To address the problems of context, in 2002, in consultation with Beth Tillinghast from the UHM Library's Desktop Network Services, the Hawaiian and Pacific Collections began using HTTrack, which would preserve websites in the digital medium on a CD, which allowed internal links to continue to be used for navigation. However, this still was not optimal, as access was limited to patrons who physically came to the library (a step backward from the original web-based item). The Internet Archive (a nonprofit digital library headquartered in San Francisco) was established in 1996 with the ambitious goal of archiving the whole web. However, Hawaiian and Pacific Collection librarians had no control over what websites were selected by the Internet Archive and how much of each website was preserved. Very short-term information like that from websites reporting on elections or natural disasters would not necessarily be preserved. In 2006, to address this need, the Internet Archive developed Archive-it, a system that allows subscribers to select which web pages are crawled and preserved by the Wayback Machine. In 2008, Dore Minatodani, librarian from the Hawaiian Collection, and Stuart Dawrs, librarian from the Pacific Collection, joined Tillinghast in launching a pilot instance of Archive-it for just this purpose.

Selecting websites also requires an additional effort on the part of the librarian. For books, the publisher decides on the size and shape of the item, and librarians purchase it and process it as is. When archiving websites, the librarian determines the size, and to some extent, the "beginning" and the "end" of the website, when defining how much of the website will be preserved. Some websites are preserved in their entirety, with all links, embedded videos, and posted PDFs accessible and intact. For other websites, for reasons of copyright or space limitations, all that is captured is the first page. The proof of existence and the look and feel of the website is preserved, but the bulk of the content is not.

The librarian must also choose how often to preserve the website. Each instance of preservation is capturing and preserving a static snapshot of what the website looked like and contained on that day. Some websites are relatively stable and are not updated often and therefore may not require frequent captures. Other sites, such as blogs and websites featuring current events, may dramatically change on a monthly, weekly, or even daily basis. While it is beyond the means of this project to capture websites on even a weekly basis, it is important to note that when capturing an instance of a website, it is only one of hundreds, or possibly even thousands of versions that are not captured and preserved.

How We Select Websites

As of 2013, the UHM Hawaiian and Pacific Collections subscription with Archive-it allows for the annual collection of over 6.5 million pages. Half of this is devoted to archiving University of Hawai'i websites, and the other half is split between the Hawaiian and Pacific Collections.[6] At our current subscription level, the Pacific Collection is able to annually preserve roughly 50 percent of the websites identified as being within the scope of this project.

The Pacific Collection collects and organizes websites from twenty-six of the different countries and territories within its regional scope. One exception is that while we do collect print material related to the Māori experience, we have chosen not to collect Aotearoa/New Zealand websites, for reasons mentioned below. At the time of this writing, the Pacific Collection has archived 667 websites and identified an additional 256 to add for the next fiscal year (2013–2014). The Hawaiian Collection collects and organizes websites within fifteen broad subjects and, to date, has archived 242 of them. Although some websites are only collected once,

most will be collected multiple times during their existence. With the continued expansion of the web, it is expected that the overall number of websites worthy of collection will increase each year.

Websites are selected according to whether they are created in the Pacific Islands, by Pacific Islanders, or about the Pacific Islands. This broad scope loosely parallels the Pacific Collection's comprehensive collection development policy and is intended to ensure a wide representation of information and experiences. In searching for potential websites to archive, active searching and serendipity both play an important role. To find websites within the web domain of a country or territory (for example, ".fm" for the Federated States of Micronesia or ".ws" for Sāmoa), it is possible to conduct an advanced Google search limited by domain code. Additionally, potential relevant websites are also often found mentioned in the same sources as print resources, such as listserves, academic and popular journals, social media, and news sources. Suggestions from website creators, owners, or users are also welcome.

Just as the Pacific Collection has made a significant effort to collect government documents in print form, the websites of Pacific Island governments are a priority for archiving. The websites of nongovernmental and regional organizations as well as national or regional associations are also considered of high importance. There are a great many science, social science, and humanities research websites in the Pacific Island region. The individual voices of Pacific Islanders are preserved in blogs and some forum discussions. An effort is also made to capture current events, such as the response to natural disasters or political change. Examples of these categories of archived sites include, respectively, Guam's Department of Public Health and Social Services;[7] the Pacific Institute of Public Policy;[8] Institut de recherche pour le développement (IRD) Nouvelle-Calédonie;[9] Planet Tonga;[10] and several sites about the 2009 earthquake and tsunami in Sāmoa.[11]

Future Directions

As this project continues and matures, there are several potential improvements and new possibilities. The first is to enhance the access and visibility of those websites that we have already archived. The Hawaiian and Pacific Collections have collaborated with the UHM Library's Cataloging Department to establish a standard for describing (cataloging) websites and how to make that description available in the Library's various search tools.

An online library guide has also been created to explain the project and facilitate access to the archived websites.[12]

This project is also being strengthened by collaborations with other collecting institutions. The National Library of New Zealand (NLNZ) created and maintains the New Zealand Web Archive.[13] The New Zealand legal deposit legislation, passed in 2006, allows the NLNZ to collect New Zealand digital publications, including websites. This legislation does not, however, cover websites created outside of New Zealand, meaning any website without the ".nz" domain. The NLNZ is still interested in documenting Pacific websites, especially those from countries and territories with historical or current relations with New Zealand.

Like the UHM Hawaiian and Pacific Collections, the National Library of Australia (NLA) also subscribes to Archive-it for the purposes of archiving websites. Australia does not yet have legal deposit legislation that applies to the digital realm, but the NLA has been archiving websites from Pacific Island countries and territories with which is it historically linked, mostly focused on websites related to elections.

In December 2012, representatives from NLNZ, NLA, and the UHM Hawaiian and Pacific Collections met to discuss a collaborative approach to archiving the Pacific web. The purpose and scope of web archiving conducted by each institution was compared, so as to limit duplication and identify any possible gaps.

Another important future direction for this project is active collaboration with those who own these websites. Currently the UHM Hawaiian and Pacific Collections rely on the permissions model of Archive-it itself, which states that if the website owner objects to the website being archived, Archive-it will remove the item from its servers. We would like to move toward a permissions model that actively engages and informs website creators of the value of the service and of their options regarding the website archiving process. This will be a long-term and ongoing effort, with over 1,000 websites already being archived at the University of Hawai'i (and the numbers continue to grow). One positive solution is to use the Pacific Collection's annual acquisitions trips as an opportunity to meet with website owners and begin the discussion face-to-face. Many website owners are concerned about external access to private web-based data, and it is important to reassure them that only publically available and visible sections of the website are archived.

In addition to website owners, we need to engage with the national or territorial libraries and archives of the Pacific region. These are the insti-

tutions responsible for documenting and preserving national or territorial history, and, in the short term, support from these institutions for the UHM Hawaiian and Pacific Collections web-archiving project would significantly help in building relations with website owners. But these institutions are also in need of support. Laws in Pacific Island countries related to the preservation of information are weak (see, eg, Cass 2008, 6), and currently none include provisions for the preservation of websites. There are a few Pacific Island countries with legal deposit legislation, which supports the national libraries or archives by requiring that print copies be supplied for preservation purposes. A longer-term goal would be to ensure that electronic publications and websites are also explicitly included in legal deposit legislation where it exists and to encourage the creation of such legislation where it does not exist.

Notes

1 Library of Congress Archived Web Sites (http://www.loc.gov/websites/collections/), accessed 28 August 2013.

2 The other international initiative is the Latin American Web Archiving Project (LAWAP), hosted by the University of Texas at Austin.

3 The UHM Library's Hawaiian and Pacific Collections are two distinct, geographically defined collections, with the Pacific Collection responsible for the Pacific Island countries and territories within Micronesia, Melanesia, and Polynesia, excluding Hawai'i.

4 Recent SSGM discussion papers can be found at http://ips.cap.anu.edu.au/ssgm/publications/recent-discussion-papers/.

5 The Pitcairn Miscellany website is available to subscribers at http://www.miscellany.pn/.

6 The East-West Center also contributes to the subscription and receives a portion of the allocation for the purpose of archiving its institutional websites.

7 Guam's Department of Public Health and Social Services is archived at http://wayback.archive-it.org/1253/*/http://dphss.guam.gov/.

8 The Pacific Institute of Public Policy is archived at http://wayback.archive-it.org/2645/*/http://www.pacificpolicy.org/.

9 The IRD website is archived at http://wayback.archive-it.org/2267/*/http://nouvelle-caledonie.ird.fr/.

10 The Planet Tonga website is archived at http://wayback.archive-it.org/2899/*/http://www.planet-tonga.com/.

11 Websites about the 2011 earthquake and tsunami in Sāmoa are archived

at http://www.archive-it.org/collections/1889;jsessionid=AE2C73D737FDB3E700049E687D09BC30/.

12 The UHM library guide can be found at http://guides.library.manoa.hawaii.edu/pacificwebarchive/.

13 The New Zealand Web Archive is at http://natlib.govt.nz/collections/a-z/new-zealand-web-archive/.

References

Cass, Libby

2008 IFAP Principles in Action in the Pacific. Paper presented at the Pacific IFAP [Information for All Programmes] Committee meeting, Wellington, New Zealand, 6 May.

Gomes, Daniel, João Miranda, and Miguel Costa

2011 A Survey on Web Archiving Initiatives. *Proceedings of the 15th International Conference on Theory and Practice of Digital Libraries,* edited by Stefan Gradmann, Francesca Borri, Carlo Meghini, and Heiko Schult, 408–420. Heidelberg: Springer-Verlag. Berlin.